Electronic Supplementary Material (ESI) for Digital Discovery.

Supplementary Information: A database of molecular properties integrated in the Materials Project

Evan Walter Clark Spotte-Smith,[*,†,‡] Orion Archer Cohen,[¶] Samuel M. Blau,[§] Jason M. Munro,[†] Ruoxi Yang,[†] Rishabh D. Guha,[†] Hetal D. Patel,[†,‡] Sudarshan Vijay,[‡,‖] Patrick Huck,[†] Ryan Kingsbury,[⊥] Matthew K. Horton,[#] and Kristin A. Persson[*,‡,@]

†*Materials Science Division, Lawrence Berkeley National Laboratory*

‡*Department of Materials Science and Engineering, University of California, Berkeley*

¶*Department of Chemistry, University of California, Berkeley*

§*Energy Storage and Distributed Resources, Lawrence Berkeley National Laboratory*

‖*Present address: VASP Software GmbH*

⊥*Department of Civil and Environmental Engineering, Princeton University*

#*Microsoft Research*

@*Molecular Foundry, Lawrence Berkeley National Laboratory*

E-mail: espottesmith@gmail.com; kapersson@lbl.gov

Ranking Levels of Theory

Throughtout the dataset construction process of MPcules, calculations (and properties based on them) are ranked based on the level of theory used. Each component of the level of theory (density functional, basis set, and solvent method) is assigned a score (Tables S1 – S3). When ranking levels of theory, the three scores are added together, and different calculations or properties are ranked based on the negative of this sum, such that the lowest score is the most favored. In case of a tie in the level of theory, the electronic energy of a calculation is used to determine the "best" calculation; the calculation with the lowest tiebreaker score (energy) is selected.

We emphasize that these scores are entirely arbitrary, though they are guided by some basic principles. In the case of density functionals, we used benchmark studies to guide our scoring. Functionals that generally perform better than others, particularly for tasks related to thermochemistry, are favored over those that perform less well. Larger basis sets are generally favored over smaller ones, at least within a given family, and more complex solvent models accounting for e.g. non-electrostatic effects are favored over simpler models.

Functional	**Score**
ωB97X-D[1]	5
ωB97X-V[2]	6
ωB97M-V[3]	7

Table S1: Scores for different density functionals included in MPcules.

Basis Set	Score
def2-SVPD[4]	2
def2-TZVPPD[4]	6
def2-QZVPPD[4]	7

Table S2: Scores for different basis sets included in MPcules.

Solvent Method	Score
Vacuum	1
PCM[5]	3
SMD[6]	5

Table S3: Scores for different solvent methods included in MPcules.

MPcules composition by level of theory

Density Functional	Basis Set	Solvent Model	Number of Molecules
ωB97X-D	def2-SVPD	Vacuum	77
ωB97X-D	def2-SVPD	PCM	95
ωB97X-D	def2-SVPD	SMD	103
ωB97X-V	def2-TZVPPD	SMD	43,041
ωB97M-V	def2-SVPD	Vacuum	102,555
ωB97M-V	def2-SVPD	PCM	2,963
ωB97M-V	def2-SVPD	SMD	33,823
ωB97M-V	def2-QZVPPD	SMD	30,871

Table S4: Number of (collected) molecules in MPcules for which calculations have been performed at various levels of theory. Note that the sum of these numbers does not equal the number of molecules in MPcules, as many molecules have been the subject of calculations at several levels of theory.

Comparison of Atomic Partial Charges and Spin

MPcules contains partial atomic charges and partial atomic spins calculated using four methods: Mulliken population analysis, the restrained electrostatic potential (RESP), Bader charges, and natural atomic chages and spins from NBO. Here, we compare the Mulliken and NBO methods, specifically focusing on the oxidation states of metals (Li and Mg). To ensure a fair comparison, we only included the charges and spins of Li and Mg atoms for which both Mulliken and NBO populations were available, and we are only comparing data points from the same solvent environment. In the case of Li (Figures S1 – S2), all calculations were performed with the SMD implicit solvent model using the parameters for a mixture of ethylene carbonate (EC) and ethyl methyl carbonate (EMC). For Mg, we compare values for two SMD solvents - diglyme (G2; Figures S3 – S4) and tetrahydrofuran (THF; Figures S5 – S6).

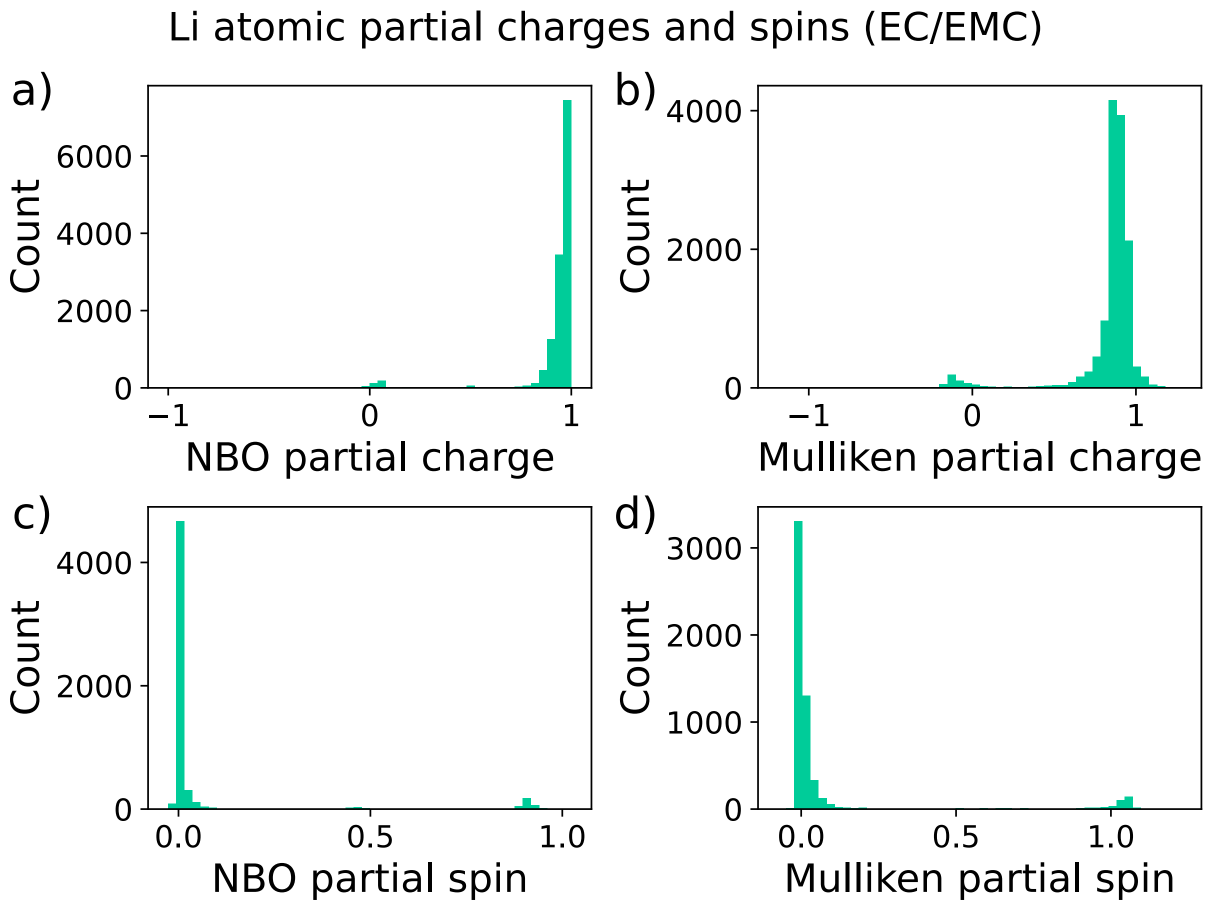

Figure S1: Histogram of Li atomic partial charges (a-b) and spins (c-d) in MPcules as calculated using the NBO (a, c) and Mulliken (b, d) methods. All calculations were performed in implicit solvent using SMD with parameters relevant for a mixture of EC and EMC.

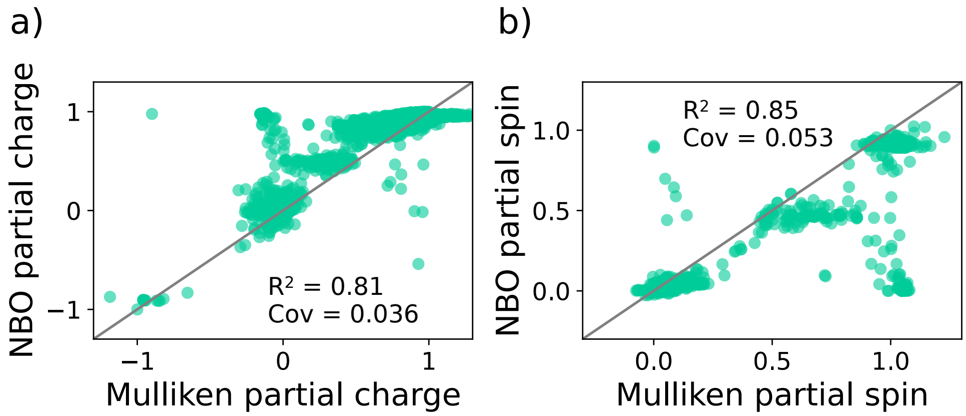

Figure S2: Comparison of Li partial atomic charges (a) and spins (b) in MPcules calculated using the NBO and Mulliken methods. All calculations were performed in implicit solvent using SMD with parameters relevant for a mixture of EC and EMC. Coefficients of determination (R^2) and covariances between the NBO and Mulliken values are provided.

In general, we find that NBO produces narrower distributions of partial atomic charges than the Mulliken method, and these distributions are centered around integral oxidation states (e.g. 0, +1 for Li and 0, +1, and +2 for Mg). For Mg in particular, the Mulliken method is often in qualitative disagreement regarding the metal oxidation state. In both G2 (Figure S3) and THF (Figure S5), most of the distribution of NBO partial atomic charges are just below 2 (indicating unreduced Mg^{2+}, while the Mulliken distribution is centered just above a charge of 1 (indicating radical Mg^{1+}).

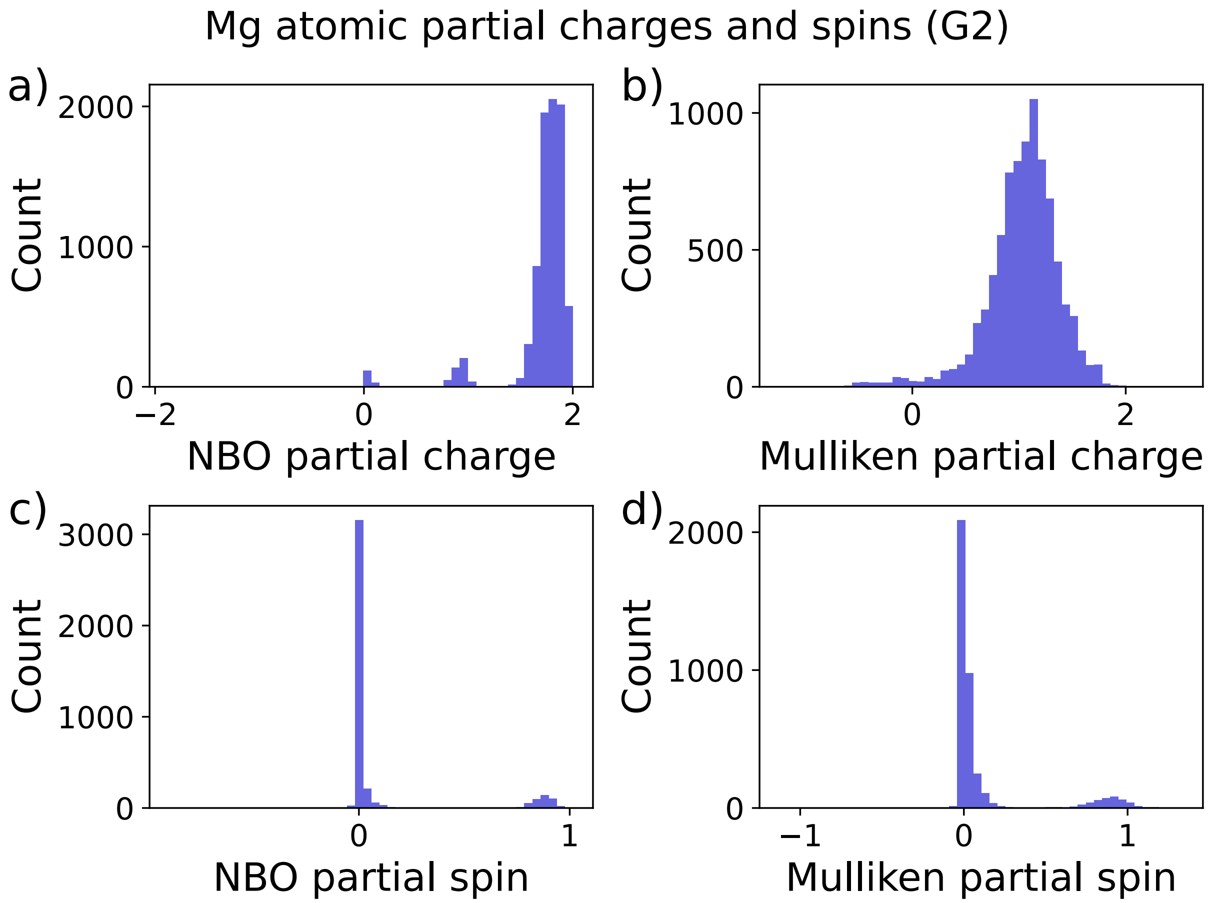

Figure S3: Histogram of Mg atomic partial charges (a-b) and spins (c-d) in MPcules as calculated using the NBO (a, c) and Mulliken (b, d) methods. All calculations were performed in implicit solvent using SMD with parameters relevant for G2.

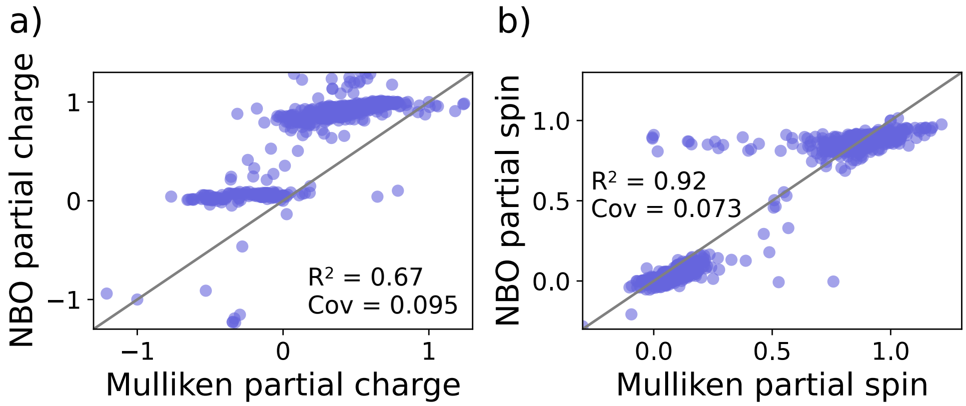

Figure S4: Comparison of Mg partial atomic charges (a) and spins (b) in MPcules calculated using the NBO and Mulliken methods. All calculations were performed in implicit solvent using SMD with parameters relevant for G2. Coefficients of determination (R^2) and covariances between the NBO and Mulliken values are provided.

In contrast to partial atomic charges, where Mulliken predictions appear to be poorly behaved compared to NBO, Mulliken partial atomic spins are in qualitative agreement with NBO. Though the distributions of Mulliken spins are still generally broader than those of NBO, both Mulliken and NBO tend to predict partial spins on Li and Mg that are close to either 0 or 1 (though there also appears to be a nontrivial number of Li atoms with partial atomic spin around 0.5). For both metals and all solvents tested, NBO and Mulliken partial atomic spins are better correlated in terms of R^2 than the corresponding partial atomic charges, further supporting the notion that Mulliken and NBO partial atomic spins are in better agreement than Mulliken and NBO partial atomic charges. From this, we tentatively suggest that partial atomic spins may be easier to capture accurately than partial atomic charges for metal atoms.

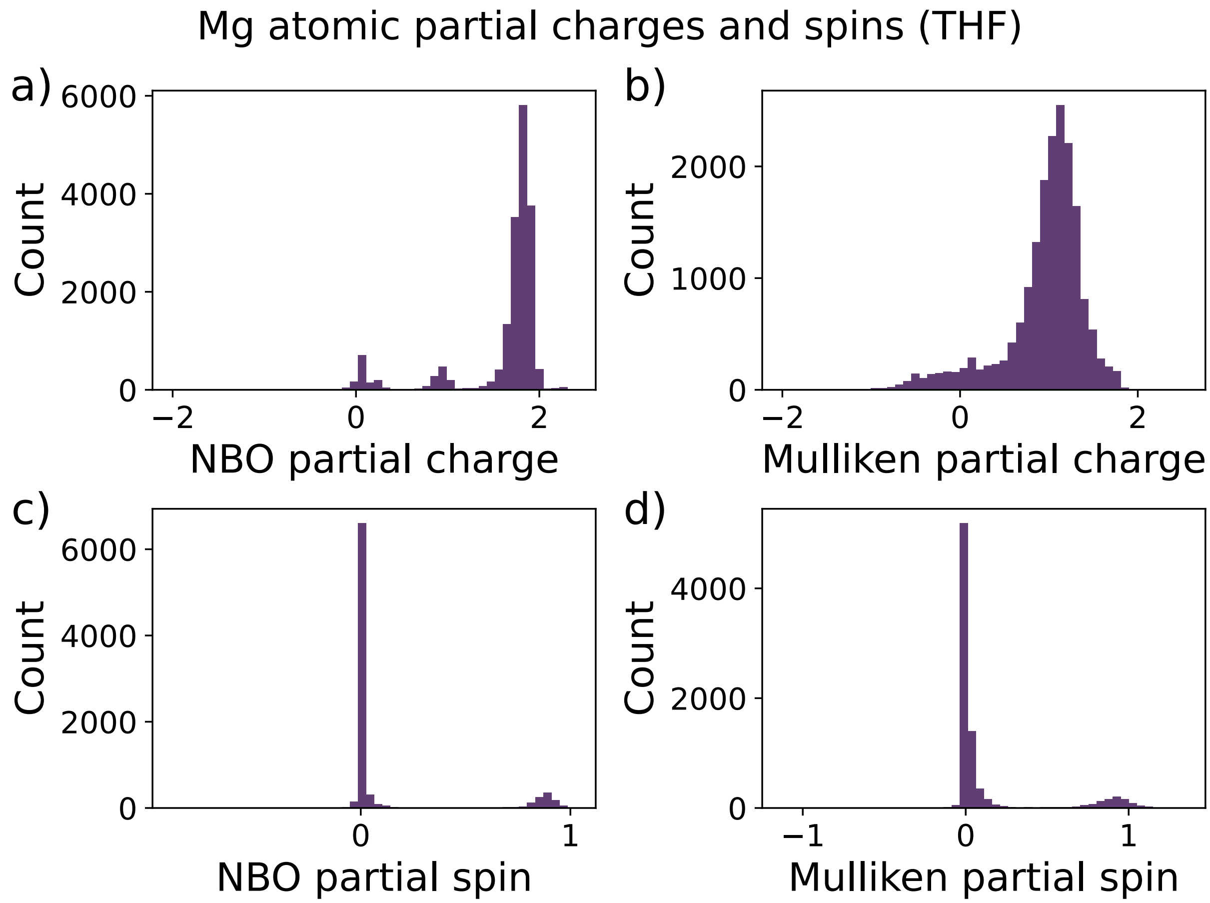

Figure S5: Histogram of Mg atomic partial charges (a-b) and spins (c-d) in MPcules as calculated using the NBO (a, c) and Mulliken (b, d) methods. All calculations were performed in implicit solvent using SMD with parameters relevant for THF.

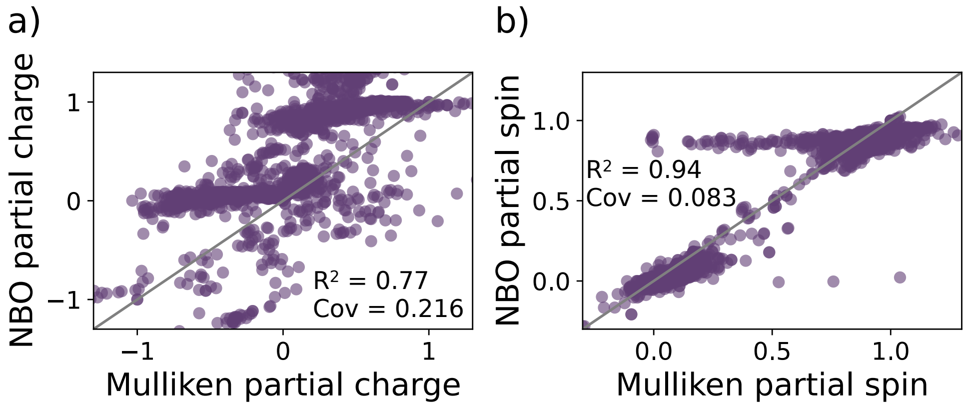

Figure S6: Comparison of Mg partial atomic charges (a) and spins (b) in MPcules calculated using the NBO and Mulliken methods. All calculations were performed in implicit solvent using SMD with parameters relevant for THF. Coefficients of determination (R^2) and covariances between the NBO and Mulliken values are provided.

References

(1) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Physical Chemistry Chemical Physics* **2008**, *10*, 6615–6620.

(2) Mardirossian, N.; Head-Gordon, M. B97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy. *Physical Chemistry Chemical Physics* **2014**, *16*, 9904–9924.

(3) Mardirossian, N.; Head-Gordon, M. B97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. *The Journal of Chemical Physics* **2016**, *144*, 214110.

(4) Rappoport, D.; Furche, F. Property-optimized Gaussian basis sets for molecular response calculations. *The Journal of Chemical Physics* **2010**, *133*, 134105.

(5) Mennucci, B. Polarizable continuum model. *WIREs Computational Molecular Science* **2012**, *2*, 386–404.

(6) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *The Journal of Physical Chemistry B* **2009**, *113*, 6378–6396.